DOCUMENT RESUME

ED 163 042 TM 008 060

AUTHOR Merrill, M. David; Wood, Norman D.
TITLE Validation of the Instructional Strategy Diagnostic Profile (ISDP): Empirical Studies. Final Report.
INSTITUTION Navy Personnel Research and Development Center, San Diego, Calif.
REPORT NO NPRDC-TR-77-25
PUB DATE Apr 77
CONTRACT N00123-76-C-0458
NOTE 65p.

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS *College Curriculum; *Curriculum Development; *Curriculum Evaluation; Evaluation Criteria; *Evaluation Methods; Higher Education; *Instructional Improvement; *Instructional Materials; Performance Factors; Physics Curriculum; Statistics; Teaching Methods

ABSTRACT

The Instructional Strategy Diagnostic Profile (ISDP) was designed to enable instructional developers and evaluators to predict the effectiveness of, and prescribe improvements for existing instructional materials. Experimental studies were conducted in introductory college statistics and physics classes to validate the ISDP and its accompanying design prescriptions. Two methodologies were used: (1) existing instructional materials were modified, based on a selected prescription resulting from ISDP analysis of those materials, and two or more versions of the materials were compared; and (2) a weak unit of an existing course was identified and modified via several prescriptions resulting from an ISDP analysis. Test performance, affect, confidence, and time were compared for students using the revised materials and students using the original materials. When used to revise existing materials, the ISDP prescriptions produced significant differences only in the second study. Failure to obtain predicted results may have been due to confounding factors in the experimental situations. Other studies have demonstrated that existing materials revised according to ISDP prescriptions can significantly improve student performance, especially if the interaction with the materials can be controlled and the tests can be revised to more adequately measure concept classification and rule-using behavior. (Author/JAC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED163042

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

NPRDC TR 77-25

April 1977

VALIDATION OF THE INSTRUCTIONAL STRATEGY DIAGNOSTIC PROFILE (ISDP): EMPIRICAL STUDIES

M. David Merrill
Norman D. Wood

Courseware, Inc.
San Diego, California 92131

Reviewed by
John D. Ford, Jr.

Approved by
James J. Regan
Technical Director

Prepared for
Navy Personnel Research and Development Center
San Diego, California 92152

2-A

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NPRDC TR 77-25	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): VALIDATION OF THE INSTRUCTIONAL STRATEGY DIAGNOSTIC PROFILE (ISDP): EMPIRICAL STUDIES	5. TYPE OF REPORT & PERIOD COVERED: Final Report
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): M. David Merrill Norman D. Wood	8. CONTRACT OR GRANT NUMBER(s): N00123-76-C-0458
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Courseware, Incorporated San Diego, California 92131	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 63720N Z0108.30A
11. CONTROLLING OFFICE NAME AND ADDRESS: Navy Personnel Research and Development Center San Diego, California 92152	12. REPORT DATE: April 1977
	13. NUMBER OF PAGES: 64
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Instructional Strategies
Statistics
Feedback
Advanced Organizers
Instructional Strategies Diagnostic Profile (ISDP)
Instructional Diagnosis
Instructional Prescription

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Three experimental studies were conducted in real world settings in an attempt to validate the Instructional Strategy Diagnostic Profile and the accompanying design prescriptions.

Two different methodologies were used. In method one, existing instructional materials were modified on the basis a selected prescription that resulted from an ISDP analysis of those materials. Two or more versions of the materials were compared in an experimental comparison. Method two

consisted of course intervention in which a weak unit of an existing course was identified and modified via several prescriptions resulting from an ISDP analysis. Test performance, affect, confidence, and time were compared for students using the revised materials and students using the original materials.

When used to revise existing materials, the ISDP prescriptions produced significant differences only in the 2nd study. Failure to find the predicted results may have been a result of confounding factors in the real-world experimental situations used. Other studies have demonstrated that existing materials revised according to ISDP prescriptions can be demonstrated to produce significant increases in student performance especially if the interaction with the materials can be controlled and the tests can be revised to more adequately measure concept classification and rule-using behavior.

FOREWORD

This research and development was conducted in support of Advanced Development Subproject Z0108.30A (Adaptive Experimental Approach to Instructional Design). This work is one aspect of an area concerned with evaluation of instruction/training. Previous work reviewed the existing research literature that has investigated the propositions underlying the Profile. The review identified requirements for research and development to which the studies described in the present report, which involved test and evaluation of the ISDP, were directed.

Drs. John Carter and John Ellis served as contract monitors. This report was reviewed and edited by Dr. John Ellis.

J. J. CLARKIN
Commanding Officer

SUMMARY

Problem

The Instructional Strategy Diagnostic Profile (ISDP) was designed to enable instructional developers and evaluators to predict the effectiveness of and prescribe improvements for existing instructional materials. While some evidence exists as to its effectiveness, it has not yet received sufficient empirical evaluation.

Objective

The purpose of this research and development effort was to further the development and validation of the ISDP and the accompanying instructional design prescriptions.

Approach

Three empirical studies were conducted using two different methodologies. Method one, which was used for the first two studies, consists of modifying existing instructional materials based on prescriptions resulting from an ISDP analysis of those materials. Method two, which was used for the third study, is an intervention process, in which a weak unit of an existing course is selected and modified via several prescriptions resulting from an ISDP analysis. Test performance, affect, confidence, and time were compared for students using the revised materials and for those using the original materials.

Study 1, using method one on an introductory statistics course, compared a framework presentation with a regular prose presentation with either elaborated or correct answer feedback.

Study 2, using method one on the same statistics course as Study 1, compared the performance of groups using connected rules, discrete rules, and rules embedded in expository text. The connected rule involved an algorithm for selecting which rule to use; the discrete rule contained an algorithm; and the embedded rule neither involved or contained an algorithm.

Study 3, using method two, revised a unit of the syllabus for an introductory physics course and then compared that unit with the original unit in terms of student performance, time, and affect. Lectures, textbooks, and discussions were the same for both groups.

Findings

In Study 1, no significant differences were observed in student performance, affect, confidence, or time. In Study 2, posttest performance of both discrete and connected rule groups was superior to that of the embedded rule group. On affect, the discrete rule was most positive, followed by the connected and embedded rule groups. There were no time or confidence differences. Finally, in Study 3, there were no significant differences in performance, time, or affect between the two groups, although the means were in the predicted direction. Failure to find the predicted results may have been a result of confounding factors in the real-world experimental situations used.

Conclusions

The research reviewed indicate that the propositions underlying the Profile seem to be valid. While the data reported in this document is somewhat inconclusive and not sufficient to make unqualified statements it is, nevertheless, positive. When considered with other data on the ISDP, it seems reasonable to assume that, when the ISDP I[illegible]file is used as a guide to analyze and modify existing instruction, the resulting performance of students is likely to be more effective. This is especially likely when the tests as well as the main line instruction can be modified. It is less likely when only the student syllabus is modified. The ISDP does seem to have considerable potential as an instructional evaluation and development tool.

Recommendations

1. The ISDP, as presented in the ISDP training manual, is recommended for use by Navy instructional developers and evaluators. However, it should be considered as an experimental tool and should be used only by experienced instructional technologists who can appropriately adapt its use to various settings and circumstances.

2. It is recommended that ISDP validation and development efforts continue so that this instrument can become an easy to use tool for all instructional development and evaluation personnel.

CONTENTS

	Page
INTRODUCTION	1
Problem	1
Purpose	1
Background and Scope	1
APPROACH	3
Study 1--Framework Rule Representation and Elaborated Feedback in Statistics Instruction	3
Study 2--Test and Generality Consistency in a Classification Task	3
Study 3--Validation of the Instructional Strategy Diagnostic Profile in Physics 100	4
STUDY 1: FRAMEWORK RULE REPRESENTATION AND ELABORATED FEEDBACK IN STATISTICS INSTRUCTION	5
Design Challenges	5
Rule Representation	5
Appropriate Practice	5
Hypotheses	6
Methods	6
Selection of Subject Matter	6
Subjects	6
Treatment Materials	7
Instrumentation	7
Design	22
Results	22
Discussion	24
STUDY 2: TEST AND GENERALITY CONSISTENCY IN A STATISTICS CLASSIFICATION TASK	25
Problem	25
Hypothesis	26
Method	27
Subject Matter Content	27
Subjects	27
Treatments	27
Instrumentation	30
Procedure	33
Design	33

Page

Results 34

Hypothesis 1 34
Hypothesis 2 34
Hypothesis 3 35
Hypothesis 4 35

Discussion 35

STUDY 3: VALIDATION OF THE INSTRUCTIONAL STRATEGY DIAGNOSTIC PROFILE IN PHYSICS 100 37

Overview and Hypotheses 37
Methods 37

Subject Matter Content 37
Subjects 38
Treatments 38
Procedure 38
Design 39

Results 39

Hypothesis 1 39
Hypothesis 2 40
Hypothesis 3 40

Discussion 40

CONCLUSIONS 43

RECOMMENDATIONS 45

REFERENCES 47

APPENDIX - EVALUATION OF TEST ITEM TYPE AND STUDENT PERFORMANCE IN PHYSICS 100 . . . . . A-0

DISTRIBUTION LIST

LIST OF TABLES

Page

1. Means and Standard Deviations of Dependent Variables for Four Treatment Groups 23
2. Summary of Univariate F-Ratios on Four Dependent Variables 24
3. Means and Standard Deviations for Dependent Variables by Treatment Group 34
4. Mean Percentage Correct and Standard Deviations by Treatment Group . 40
5. Mean Percentage of Items Answered Correctly for Students with Upgraded and Regular Materials 41

LIST OF FIGURES

1. Framework rule representation with elaborated feedback 8
2. Framework rule representation with correct answer feedback 11
3. Nonframework rule representation with elaborated feedback 14
4. Nonframework rule representation with correct answer feedback . . . 17
5. Dependent variable measures including an example of a rule-using test item, a confidence scale, a semantic differential affect scale, and a time record space 20
6. Connected-rule algorithm for selection of hypothesis test 28
7. A sample practice question 29
8. An example of a discrete rule algorithm for selection of hypothesis test 31
9. A sample of the basic performance task and confidence rating 32

INTRODUCTION

Problem

Guidelines for predicting instructional effectiveness, if they exist at all, are vague at best. It is almost impossible to look at an instructional product and predict its effectiveness by the use of existing guides. The Instructional Strategy Diagnostic Profile (ISDP) was designed (Merrill & Wood, 1975) to enable instructional developers and evaluators to predict the effectiveness of and prescribe improvements for existing instructional materials. While some evidence exists as to its effectiveness, it has not yet received sufficient empirical validation.

Purpose

The purpose of this research and development effort was to further the development and validation of the ISDP and the accompanying instructional design prescriptions.

Background and Scope

This project was conducted in three phases. Phase I consisted of an extensive review of reported research studies as they relate to the propositions underlying the Instructional Strategy Diagnostic Profile (Merrill, Olson, & Coldeway, 1976). Results of this review indicate that there is considerable empirical research support for most of the propositions underlying the Profile. It was further suggested that an instructional package that is judged to have a high ISDP index should provide rather effective instruction.

This document is the technical report for the Phase II effort, which consisted of three empirical studies using two different methodologies.

Phase III involved the preparation and validation of a manual for training users in ISDP analysis. This manual will be published as a separate technical report (Merrill, Wood, & Richards, in preparation). Preliminary validation indicates that experienced instructional developers, who have already had some training in the vocabulary of the ISDP, are able to consistently rate existing instruction and to prescribe modifications by using the guidelines provided by the ISDP training manual.

APPROACH

As indicated previously, Phase II of this project consisted of conducting three empirical studies using two different methodologies. Method one, which was used for the first two studies, consists of modifying existing instructional materials based on prescriptions resulting from an ISDP analysis of those materials. Method two, which was used for the third study, is an intervention process, in which (1) a course is analyzed via the ISDP, (2) a weak unit of instruction is selected, (3) the test used for that unit is revised such that it yields a higher ISDP index, and (4) the unit of instruction selected is modified so that the strategy used yields a higher ISDP index. The original and revised instructions are then administered to randomly assigned groups, and performance on both is compared. The three studies are described briefly below and in detail in the following sections.

Study 1--Framework Rule Representation and Elaborated Feedback in Statistics Instruction

Using method one in an introductory statistics course, a framework rule representation was compared with a regular prose rule representation and correct answer feedback was compared with elaborated feedback. On the posttest, there were no performance differences. Students were also compared on (1) the appeal of the instruction, as measured by a questionnaire, (2) confidence in their responses, and (3) time required to complete the instruction. There were no significant differences on any of these dependent measures. The ISDP rating of the instruction before the modification was very good. It was suggested that the relationship between performance and the ISDP index is a decreasing function. That is, it requires a large increment at the high end of the ISDP scale to result in a measurable performance difference, as compared to a small increment at the low end of the scale.

Study 2--Test and Generality Consistency in a Classification Task

Also using method one in the same introductory statistics course, performance of groups using three types of rule statements were compared: connected rules, discrete rules, and the regular expository test, which served as the control. The first two treatments consisted of an integrated algorithmic flow chart representation for the connected condition and separate nonintegrated flow chart representations for the discrete condition. The regular condition did not present a how-to-use-the-rule algorithm of any kind. On posttest performance, the connected and discrete rule groups scored significantly higher than the regular instruction group. On affect, all three groups were different with the discrete group most positive, the connected next, and the regular instruction least positive. There were no significant differences between groups on posttest time or response confidence. It was concluded that providing the student with an algorithmic rule resulted in a performance increment. It should be noted that this study also used materials which had a high initial ISDP rating and that the addition of an algorithmic representation of the rule was still able to cause a further increment in performance.

Study 3--Validation of the Instructional Strategy Diagnostic Profile in Physics 100.

The second methodology was used in an introductory physics course at Brigham Young University. An ISDP analysis of the tests used in the course indicated that less than 20 percent of the test items met the ISDP requirements for adequate rule-using. Performance on the rule-using items was significantly lower than performance on the memory-oriented items. One of the poorest (as indicated by test performance) units of instruction was selected for ISDP modification. A revised test was prepared which included more rule-using items, and the instruction was revised to score higher on the Profile. During the summer term, students in the course were randomly assigned to the existing or the revised materials. All students took both the original and the revised test.

There was a significant difference within groups, indicating that performance on encountered test items was better than that on unencountered test items. However, between-group differences, on either type of test, while in the predicted direction, failed to reach significance. The study attempted to demonstrate that modification of the syllabus in accordance with ISDP principles would result in a performance increment. Because of considerable within group variance probably resulting from the uncontrolled influence of lectures, the textbook, and student interaction, the results failed to demonstrate the predicted difference.

STUDY 1: FRAMEWORK RULE REPRESENTATION AND ELABORATED FEEDBACK IN STATISTICS INSTRUCTION[1]

Design Challenges

The development of instruction that effectively prepares learners to use complex rules under low prompt testing conditions presents two challenges to the instructional designer: (1) finding an appropriate representation of the rule(s) that facilitates recall, and (2) providing for appropriate practice and feedback that effectively prepares the learner to apply the rules on similar test items.

Rule Representation

Several sources provide evidence for the necessity of representing rules in instruction with accompanying mathemagenic information. Landa (1974) found that the effectiveness and efficiency of student performance increased when algorithmic representation was used in teaching mathematics and language rules. Markle (1975) found that a vertical list of attributes was superior to a paragraph with embedded attributes in the acquisition of rules. Mayer (1975a) concluded that an assimilative set or framework facilitated storage and retrieval of rules from memory. Minsky (1974) suggests that a framework of information facilitates the manipulation of critical elements when encountering a new situation or instance. Glaser (1976) advocates finding ways to represent complex information to the novice learner in such ways that encoding is facilitated and time to criterion competency is decreased.

Appropriate Practice

Practice can be defined as an instructional display that (1) requires the learner to respond overtly to an explicitly stated task and (2) provides, at least, correct answer feedback to the learner. Practice is judged to be appropriate if the task, content, and feedback of the instructional display are isomorphic to the task and content of the rule representation. It is assumed in this study that the most effective practice displays should include framework displays identical to the rule representation, and that feedback should be the correct answer with elaboration rather than the correct answer only. Merrill and Wood (1975),[2] Wood, Richards, and Merrill (1976), and Schmidt, Wood, and Merrill (1976) provide a rationale and some evidence for creating example and practice displays that are isomorphic to or consistent with rule displays.

---

[1]Study conducted by N. D. Wood, R. M. Gilstrop, and M. D. Merrill.

[2]A more extensive definition of the terms used in this section is found in Merrill and Wood (1975). Space limitations do not allow for a full elaboration here.

Hypotheses

This study will investigate the effects of (1) representing complex rules within a mathemagenic framework of information and (2) providing elaborated feedback to practice displays. The general hypothesis is as follows:

> The framework (mathemagenic) representation of rules and consistent practice with elaborated feedback will produce significantly more positive student outcomes than straight list or nonframework representation of rules with correct answer only feedback.

Methods

Selection of Subject Matter

Glaser and Resnick (1972) have identified the need to do instructional psychology research within realistic settings with existing curricula. Based on this perceived need to go beyond the artificial context and subject matter of the laboratory setting, an ongoing introductory statistics course at Brigham Young University was selected for the following reasons: (1) typical subject matter within a typical instructional setting was available, and (2) complex subject matter conducive to framework rule representation was an integral part of the course.

A segment or lesson of instruction from a unit of hypothesis testing for one mean was chosen as the specific subject matter of the study. This segment included the following six steps:

1. Formulate the null (H_o) and alternative (H_a) hypotheses.
2. Choose sample size (n) and alpha.
3. Choose test statistic.
4. Make decision rule.
5. Calculate test statistic.
6. Make decision.

Subjects

A group of 93 students (encouraged by the statistics department to participate) was given course credit for participation in the experiment. Most of the students were sophomores and juniors, with a few seniors and graduate students. A wide diversity of majors was represented by the group.

Treatment Materials

The four types of treatment materials were:

1. Framework rule representation with elaborated feedback.

2. Framework rule representation with correct answer feedback.

3. Nonframework rule representation with elaborated feedback.

4. Nonframework rule representation with correct answer feedback.

All four treatment material types were in workbook form and were randomly distributed to students in the lecture hall where the experiment was conducted.

Representative examples of the rules, practice, and feedback displays used in the four treatments are found in Figures 1 through 4. Figures 1 and 2 show the framework rule representation with elaborated and with correct answer feedback, respectively; and Figures 3 and 4, the nonframework rule representation with elaborated and with correct answer feedback.

The treatment condition shown in Figure 4 represents the original instructional materials used in the course. These materials appeared in the form of a self-instructional text by Christensen (1974). An agreement was made with the instructor that any treatment that was considered by means of the Instructional Strategy Diagnostic Profile (ISDP) (Merrill & Wood, 1975) as less effective than the original would not be used. Therefore, the treatments shown in Figures 2 and 3 were considered to be more effective, while the treatment in Figure 1 was considered to be the most effective.

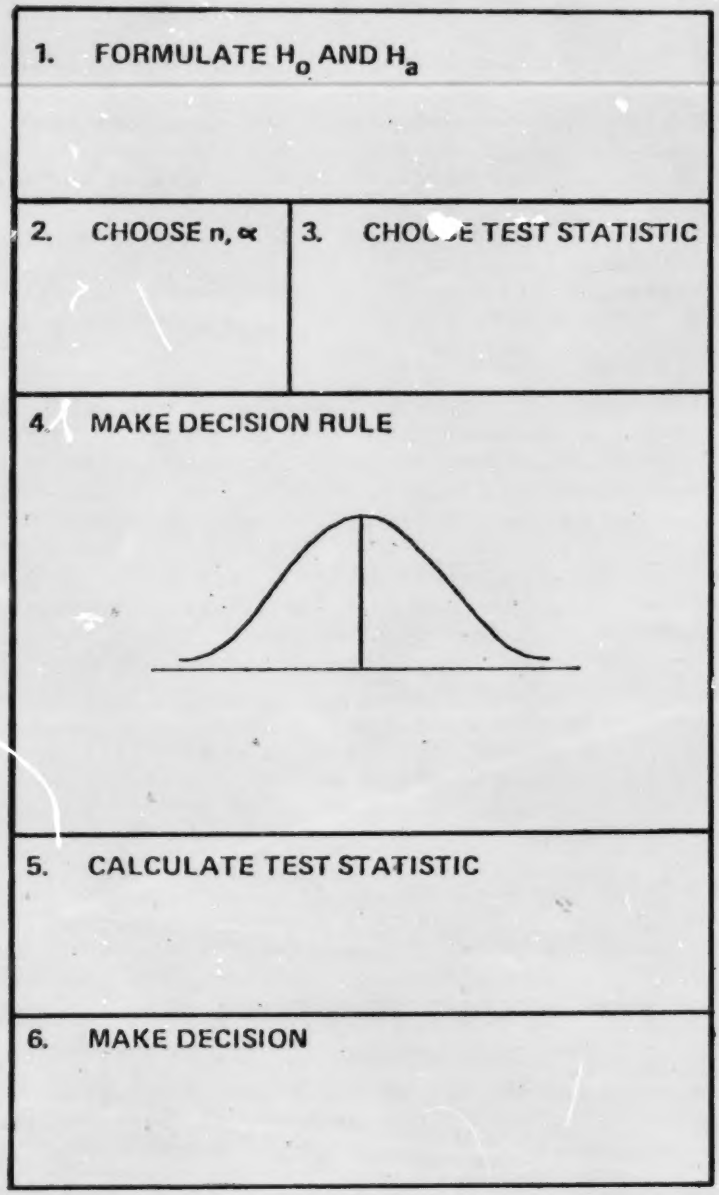
Instrumentation

The post-treatment test administered to all students in the experiment comprised 12 paper-and-pencil, multiple-choice questions of recall, concept classification, and rule-using types. An example of a rule-using test item is found in Figure 5. The number of correct answers on the twelve items by each student was used as the dependent variable performance.

Each question was followed by a seven-point differential scale, which probed the student's confidence in his answer to the test question. An example of the confidence scale item is shown in Figure 5. The average confidence for each student on all items was used as the dependent variable confidence.

The affect that the instruction had on the student was measured five times during the treatment period, using a seven-question semantic differential scale format (see Figure 5). The measures were taken after students (1) read the instructional materials, (2) responded to test items 1 and 2 (memory type), (3) responded to test items 3, 4, 5, and 6 (concept and rule using), (4) responded to test items 7, 8, 9, and 10 (concept and rule using), and (5) responded to test items 11 and 12 (higher order rule using). The word pairs were scrambled as to order and polarity for the five measures in order to avoid an anticipation effect. The average score for each student on all five affect measures was used as the dependent variable affect.

The six steps of this procedure can be remembered more easily if they are listed in a framework similar to the one below.



These six steps will be discussed in more detail within the above framework. The examples for illustration and the practice examples will be presented using this framework in order to assist in learning and remembering the 6 basic components of hypothesis testing.

Figure 1. Framework rule representation with elaborated feedback.

PRACTICE

In the following problems, use the six-step procedure discussed in this section to test the hypothesis for means of normally distributed populations when σ_x is not known:

1. The following are measurements of Brix degrees on molasses: 82.0, 79.6, 78.4, 81.8, 82.2, 79.9, 83.2 79.9, 82.3, 84.1. In order to be graded as high quality molasses, the Brix degrees must be equal to 80. At an α value of 0.05, could the molasses from which the samples were taken be graded as high quality? For this data, $\pm$ = 81.34, s = 1.8, and $\sqrt{10}$ = 3.16.

1.	
2.	3.
4.	
5.	
6.	

Figure 1. (Continued).

Answers to Practice Problem 1

1. FORMULATE H_o AND H_a

H_o: $\mu = 80$ H_a: $\mu \neq 80$

Remember that a statement of equality (either $\leq$, $\geq$, or =) will always appear in the H_o. An = sign in the H_o always defines a two-tailed test. The H_a is always the complement of the H_o.

2. CHOOSE n, α

$n = 10$

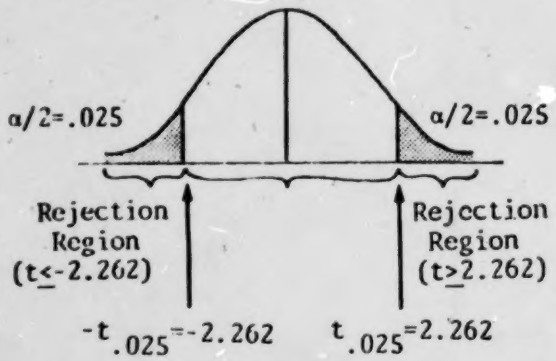
$\alpha = .05$

3. CHOOSE TEST STATISTIC

$$t = \frac{\bar{x} - 80}{s/\sqrt{10}}$$

In Step 2, α is given and n is obtained by counting the number of observations.
In Step 3, $\mu_o = 80$ and $n = 10$ are substituted into the test statistic ($t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$).

4. MAKE DECISION RULE



The $|t|$ (2.262) was obtained as follows: note: $|t|$ stands for the absolute value of t.

a. For a two-tailed test divide α by 2 ($\alpha/2 = .025$)
b. df (degrees of freedom) = $n-1 = 9$
c. Obtain t-value from table using df = 9 and $\alpha = .025$

5. CALCULATE TEST STATISTIC

$$t = \frac{81.34 - 80.00}{1.8/\sqrt{10}} = 2.354$$

Looking at the formula from Step 3, we see that the values for $\bar{x}$ and s are missing. Using the values given in the problem, ($\bar{x} = 81.34$ and $s = 1.8$) we calculate the t-value.

6. MAKE DECISION

Since $t > t_{.025}$ ($=2.354 > 2.262$) we reject the H_o.

Looking at the diagram in Step 4, we see that since $2.354 > 2.262$, we are in the rejection region of the right-hand tail. We can conclude, therfore, that the molasses cannot be graded as high quality.

Figure 1. (Continued).

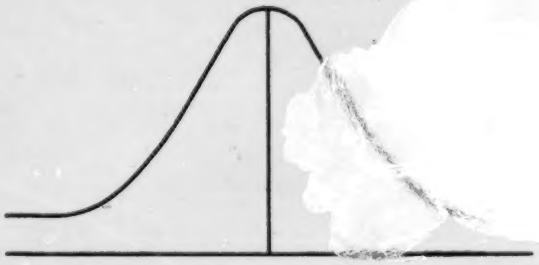
The six steps of this procedure can be remembered more easily if they are listed in a framework similar to the one below.

1. FORMULATE H_o AND H_a

2. CHOOSE n, α

3. CHOOSE TEST STATISTIC

4. MAKE DECISION RULE



5. CALCULATE TEST STATISTIC

6. MAKE DECISION

These six steps will be discussed in more detail within the above framework. The examples for illustration and the practice examples will be presented using this framework in order to assist in learning and remembering the 6 basic components of hypothesis testing.

Figure 2. Framework rule representation with correct answer feedback.

PRACTICE

In the following problems, use the 6-step procedure discussed in this section to test the hypothesis for means of normally distributed populations when σ_X is not known:

(1) The following are measurements of Brix degrees on molasses: 82.0, 79.6, 78.4, 81.8, 82.2, 79.9, 83.2, 79.9, 82.3, 84.1. In order to be graded as high quality molasses, the Brix degrees must be equal to 80. At an α value of 0.05, could the molasses from which the samples were taken be graded as high quality? For this data, $\bar{x} = 81.34$, $s = 1.8$, and $\sqrt{10} = 3.16$.

1.

2. 3.

4.

5.

6.

Figure 2. (Continued).

Answers to Practice Problems

Lesson 2 Section 2

1. (1) $H_o: \mu = 80$ $H_a: \mu \neq 80$

(2) $\alpha = .05$, $n = 10$

(3) $t = \dfrac{\bar{x} - 80}{s / \sqrt{10}}$

(4) Reject H_o if $|t| \geq 2.262$, otherwise accept H_o (note $|t|$ stands for the absolute value of t).

(5) $\bar{x} = 81.34$ $t = 2.354$

(6) Since $2.354 > 2.262$ reject H_o.

Figure 2. (Continued).

DEFINITION

A Test for One Mean When σ_x Is Not Known. A test of hypothesis for one mean when σ_x is not known is a statistical procedure used to decide whether or not the mean of a normally distributed population takes on the value of μ_o. This procedure differs from the one set forth in the previous section in the test statistic used and in the decision rules employed. In this section s is used as an estimator of σ_x. The 6 steps of the procedure are as follows:

1. Formulate H_o and H_a. The 3 possible hypotheses for the mean of a normally distributed population when σ_x is not known are:

   a. H_o: $\mu \leq \mu_o$ vs. H_a: $\mu > \mu_o$.
   b. H_o: $\mu \geq \mu_o$ vs. H_a: $\mu < \mu_o$.
   c. H_o: $\mu = \mu_o$ vs. H_a: $\mu \neq \mu_o$.

2. Choose a sample size, n, and a value for α.

3. Let the test statistic be

$$t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$$

4. On the basis of the α value, choose the decision rule according to the decision rule table, table 17.

5. Take the sample, and compute the test statistic.

6. Apply the decision rule, and make the decision.

Figure 3. Nonframework rule representation with elaborated feedback.

PRACTICE

In the following problems, use the six-step procedure discussed in this section to test the hypothesis for means of normally distributed populations when σ_x is not known:

1. The following are measurements of Brix degrees on molasses: 82.0, 79.6, 78.4, 81.8, 82.2, 79.9, 83.2 79.9, 82.3, 84.1. In order to be graded as high quality molasses, the Brix degrees must be equal to 80. At an α value of 0.05, could the molasses from which the samples were taken be graded as high quality? For this data, $\pm$ = 81.34, s = 1.8, and $\sqrt{10}$ = 3.16.

(1)

(2)

(3)

(4)

(5)

(6)

Figure 3. (Continued).

Answers to Practice Problem 1

(1) H_o: $\mu = 80$ H_a: $\mu \neq 80$

Remember that a statement of equality (either $\leq$, $\geq$, or, $=$) will always appear in the H_o. An $=$ sign in the H_o always defines a two-tailed test. The H_a is always the complement of the H_o.

(2) $\alpha = .05$ $n = 10$

α is given and n is obtained by counting the number of observations.

(3) $t = \frac{\bar{x} - 80}{s/\sqrt{10}}$

$u_o = 80$ and $n = 10$ are substituted into the test statistic $(t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}})$.

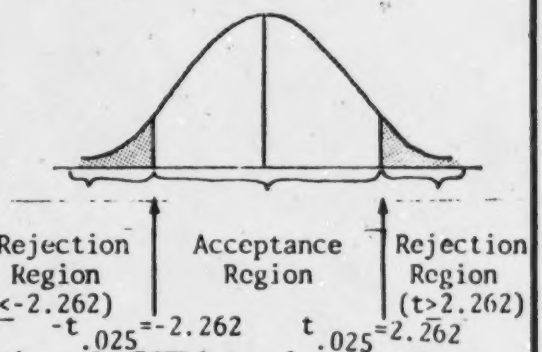
(4) Reject Ho if $|t| \geq 2.262$, otherwise accept H_o (note: $|t|$ stands for the absolute value of t)

The $|t|$ (2.262) was obtained as follows:

a. For a two-tailed test divide α by 2 ($\alpha/2 = .025$)

b. df (degrees of freedom) = n-1 = 9

c. Obtain t-value from table using df = 9 and $\alpha = .025$



(5) $\bar{x} = 81.34$ $t = 2.354$

To obtain the calculated t-value, it is necessary to look at the formula from Step 3 $(t = \frac{\bar{x} - 80}{s/\sqrt{10}})$. Using the values of $\bar{x}$ and s that were given in the problem, we substitute and compute as follows:

$$t = \frac{81.34 - 80.00}{1.8 / \sqrt{10}}$$

$$t = \frac{1.34}{.569}$$

$$t = 2.354$$

(6) Since $2.354 > 2.262$ reject H_o

Looking at the diagram in Step 4, we see that since $2.354 > 2.262$, we are in the rejection region of the right hand tail. We can conclude, therefore, that the molasses cannot be graded as high quality.

Figure 3. (Continued).

DEFINITION

A Test for One Mean When σ_χ Is Not Known. A test of hypothesis for one mean when σ_χ is not known is a statistical procedure used to decide whether or not the mean of a normally distributed population takes on the value of μ_o. This procedure differs from the one set forth in the previous section in the test statistic used and in the decision rules employed. In this section s is used as an estimator of σ_χ. The six steps of the procedure are as follows:

1. Formulate H_o and H_a. The three possible hypotheses for the mean of a normally distributed population when σ_χ is not known are:

   a. H_o: $\mu \leq \mu_o$ vs. H_a: $\mu > \mu_o$.
   b. H_o: $\mu \geq \mu_o$ vs. H_a: $\mu < \mu_o$.
   c. H_o: $\mu = \mu_o$ vs. H_a: $\mu \neq \mu_o$.

2. Choose a sample size, n, and a value for α.

3. Let the test statistic be
   $$t = \frac{\bar{X} - \mu_o}{s/\sqrt{n}}$$

4. On the basis of the α value, choose the decision rule according to the decision rule table, table 17.

5. Take the sample, and compute the test statistic.

6. Apply the decision rule, and make the decision.

Figure 4. Nonframework rule representation with correct answer feedback.

PRACTICE

In the following problems, use the 6-step procedure discussed in this section to test the hypothesis for means of normally distributed populations when σ_x is not known:

1. The following are measurements of Brix degrees on molasses: 82.0, 79.6, 78.4, 81.8, 82.2, 79.9, 83.2, 79.9, 82.3, 84.1. In order to be graded as high quality molasses, the Brix degrees must be equal to 80. At an α value of 0.05, could the molasses from which the samples were taken be graded as high quality? For this data, $\bar{x} = 81.34$, $s = 1.8$, and $\sqrt{10} = 3.16$.

(1)

(2)

(3)

(4)

(5)

(6)

Figure 4. (Continued).

Answers to Practice Problems

Lesson 2 Section 2

1. (1) $H_o: \mu = 80$ $H_a: \mu \neq 80$

   (2) $\alpha = .05$, $\eta = 10$

   (3) $t = \dfrac{\bar{x} - 80}{s / \sqrt{10}}$

   (4) Reject H_o if $|t| \geq 2.262$, otherwise accept H_o (note $|t|$ stands for the absolute value of t).

   (5) $\bar{x} = 81.34$ $t = 2.354$

   (6) Since $2.354 > 2.262$ reject H_o.

Figure 4. (Continued).

The effect of workers using new tools on the number of circuit boards assembled in an electronics plant is being tested. A random sample of individual workers' production totals is taken: 1, 3, 4, 6, 5, 5, 7, 8, 2, 2, 4, 5, 7, 3, 6, 4. The average number of circuit boards for this sample is 4.5 with a standard deviation of 2. The plant manager wants to know if the sample average of 4.5 is statistically different from the previous average of 5.5. Let $\alpha = .05$

(Use this area for work space.)

(Mark an X in the box that corresponds to the best answer for each test item below.)

7. The appropriate formulation of H_o and H_a for the above problem is:

   a. $H_o: \mu \leq 5.5$ vs. $H_a: \mu > 5.5$

   b. $H_o: \mu = 5.5$ vs. $H_a: \mu \neq 5.5$

   c. $H_o: \mu \leq 4.5$ vs. $H_o: \mu > 4.5$

   d. $H_o: \mu \geq 4.5$ vs. $H_a: \mu < 4.5$

   e. $H_o: \mu = 4.5$ vs. $H_a: \mu \neq 4.5$

   f. $H_o: \mu \geq 5.5$ vs. $H_a: \mu < 5.5$

How confident are you in your answer to the above question?

Very confident____:____:____:____:____:____:____ Not at all confident

Figure 5. Dependent variable measures including an example of a rule-using test item, a confidence scale, a semantic differential affect scale, and a time record space.

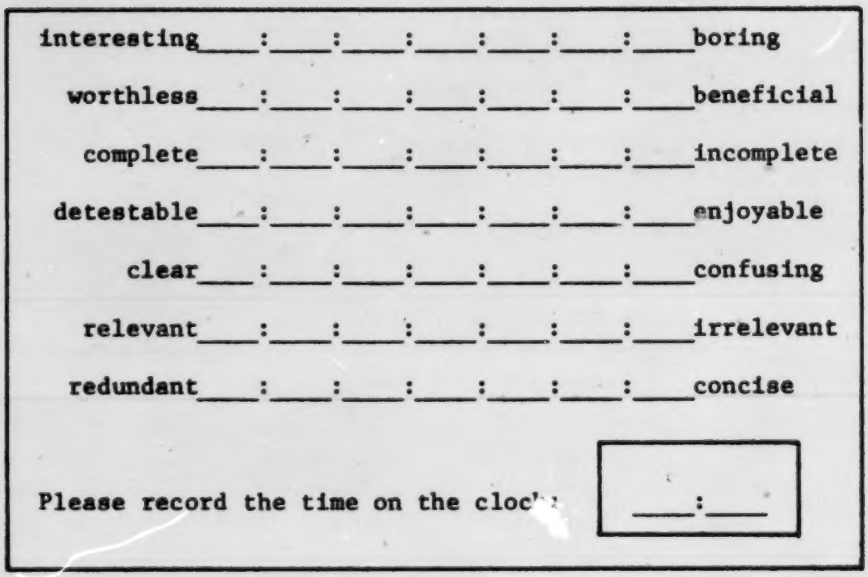


Figure 5. (Continued).

Measures of the amount of time spent by individual students on sections of the treatment materials were taken at the same five points as the affect measures. Each student wrote down the time from the wall clock in the space provided (see Figure 5). The total elapsed time taken during the treatment period by each student was used as the dependent variable time.

Design

A 2 x 2 factorial design with a multivariate analysis of variance (ANOVA) was used to assess the effects of treatments across all four dependent variables. A generalized ANOVA program which adjusted for unequal cell sizes (Bryce & Carter, 1974) was used to analyze the data.

Results

The following four hypotheses were tested:

1. Hypothesis 1

Performance scores will be higher for the framework rule/elaborated feedback treatment group than for the nonframework/correct answer treatment group.

2. Hypothesis 2

Confidence scores will be higher for the framework rule/elaborated feedback treatment group than for the nonfeedback/correct answer treatment group.

3. Hypothesis 3

Affect scores will be higher for the framework rule/elaborated feedback treatment group than for the nonframework/correct answer treatment group.

4. Hypothesis 4

Total elapsed time will be less for the framework rule/elaborated feedback treatment group than for the nonframework/correct answer feedback treatment group.

A multivariate analysis of variance (MANOVA) was performed simultaneously on all four dependent variables (performance, confidence, affect, time) as a control for an increase in Type I error through repeated univariate tests. The MANOVA F-test for the full 2 x 2 factorial model was not significant: F (4,86) = 0.697, $p > .05$. The means and standard deviations are reported in Table 1, and the respective F ratios, on Table 2.

Table 1

Means and Standard Deviations of Dependent Variables for Four Treatment Groups

Treatment Group	Dependent Variable: Performance Mean	Performance Std. Dev.	Affect Mean	Affect Std. Dev.	Confidence Mean	Confidence Std. Dev.	Time Mean	Time Std. Dev.
Framework Rule Representation:								
• Elaborated Feedback	7.25	.47	5.26	.18	4.95	.16	71.21	2.89
• Correct Answer Feedback	8.22	.48	5.21	.19	5.05	.16	72.04	2.95
Nonframework Representation:								
• Elaborated Feedback	7.39	.48	4.88	.19	4.81	.16	76.09	2.95
• Correct Answer Feedback	7.57	.48	5.41	.19	4.86	.16	74.87	2.95

Table 2

Summary of Univariate F-Ratios on Four Dependent Variables

Source of Variation	Dependent Variable			
	Performance	Confidence	Affect	Time
Rule Representation	.26	1.06	.25	1.73
Feedback	1.44	.23	1.64	.003
R x F	.69	.03	2.34	.122

Note. All F-ratios were based on df = 1,89 and α = .05.

Discussion

The mean scores on performance, confidence, and affect, as well as the total elapsed time for treatment, were chosen as the level of measurement, since more precise analysis (breaking each variable out into smaller categories) yielded no additional information.

The consistent lack of significant differences across the design in the study may be due to one or more of the following reasons:

1. The original version of the instruction (Figure 4) was considered, by means of an ISDP analysis, to be superior to any other available printed-format material on the subject. It is assumed that a less effective treatment (e.g., embedded rules in text, partial or no procedural helps for using the rule) would have assisted in creating differences between groups; in other words, by definition, the treatments were very similar.

2. The 2 hours allowed to the experimenters for the treatment period was judged to be insufficient for the complexity of the subject matter involved. The amount of information to be processed was probably too much for students, regardless of the treatment condition. It is assumed that the net effect of this time constraint drastically reduced the between-group variance that otherwise might have existed.

Additional research efforts might (1) create greater differences in treatments by embedding (or making less mathemagenic) critical attributes, and (2) allow for more time on task to determine if between-group variance can be increased.

If complex rules can be represented to the learner in ways that will develop skills of competent recall and use (application) in realistic test situations, a valuable tool for the instructional developer to increase the effectiveness and reduce the cost of instruction could be made available.

STUDY 2: TEST AND GENERALITY CONSISTENCY IN A STATISTICS CLASSIFICATION TASK[3]

Problem

Analyses of tests often indicate a required test performance that is not entirely consistent with the associated instruction. The assumption of this study is that instruction should present the student with both the content of and the behavior required for performance on a subsequent test.

One of the components of test-instruction consistency is a congruence between the test and the generality (statement of rule, definition, or proposition upon which the instruction is centered). Though some evidence exists to indicate that a generality impacts positively upon performance (Merrill, Olsen, & Coldeway, 1976), the effect of test-generality isomophism has apparently not been specifically tested.

A long time ago, Yum (1931) found that a slight change in stimulus properties from instruction to test resulted in a significant decrement of successful responses on test performance. Researchers have been slow in extending this sort of tightly controlled paired-associate study into the more complex levels of instructional application (Glaser & Resnick, 1972). At least part of the reason for this slow pace was summarized by Stake (1973), who stated that neither scales nor grounds have been developed for describing test and instruction similarity, though he cites some progress being made (e.g., Anderson, Goldburg, & Hidde, 1971). Anderson (1972) recognizes the problem in a different way when he suggests that achievement tests are based on "things" not clearly and consistently defined. Gropper (1970) has made some inroads, indicating an influence of spatial organization of materials upon student response. Mayer (1975b) has noted a forward processing effect that shows a relationahip between the kind of stimulus materials used in instruction and the test response.

Scandura's use of the algorithm and higher- and lower-order rules in instruction (Ehrenpreis & Scandura, 1974; Scandura, 1970, 1973, 1974) stresses the importance of specifying the precise behaviors requested of the learner. Shoemaker (1975) echoes this when he speaks for having identical elements in both instruction and test items. Gropper (1976) takes the position that task and content post-instructional test analysis should include the same taxonomic categories as the "front-end" instruction to effectively diagnose learning failures.

Landa (1974) concludes that students have difficulties in solving unencountered examples because the general rules necessary for identifying specific solution rules are unidentified and not taught. When this inconsistency is resolved, the integration of separate rules is facilitated, and errors decrease rapidly over a relatively short period of subsequent instruction.

---

[3]Study conducted by R. V. Schmidt, N. D. Wood, and M. D. Merrill.

All the forementioned studies point to a felt need--and some evidence--that what is tested should have been presented previously to the student (though the specific instances should differ). Just how close this match should be is open to question. A study by Scandura and Durnin (1968), indicates that a minor shift away from test-generality isomorphism is perhaps desirable.

Merrill and Wood (1974; 1975), in their Instructional Strategy Diagnostic Profile (ISDP), have taken up Stake's challenge. They have provided scales and are continuing to establish grounds for describing and evaluating concept-level instructional materials. Wood, Richards, and Merrill (1976) have developed and validated a measure of test-instruction similarity with selected constructs from the ISDP.

This study deals specifically with an assumption made in the ISDP that the test and generality should be consistent. It compares performance of students given generalities that differ in three ways in the degree to which they are consistent with the content and behaviors requested by the test items. First, a generality can present the student with content without presenting the precise task behaviors he will be asked to perform. (This does not mean that no required behavior is taught or implied, but that the specific behavior required is not taught.) This is a low level of consistency. Second, a generality can present the task conditions under which the student will be asked to perform, introducing separate generalities for each task making up a larger task. We call this "discrete rule" consistency, which requires that the mode of behavior be consistent (recall tested with recall, rule-using tested with rule-using tasks). Thus, a student is working with consistent discrete rules when an item of information he is asked to learn is taught and tested in recall mode or when a concept he is asked to learn is taught and tested with rule-using behaviors. If the item is taught with a rule-using behavior and tested in a recall mode, the test and instruction are inconsistent. The third way in which we looked at generality-test consistency involves task sequencing. This is "connected generality" consistency. Gagné (1970) and Mechner (1967) discuss this level when they describe "behavior chains." A test that asks the student to perform sequential discrete tasks in a way which is not presented in rule or practice form lacks what may be an important consistency characteristic.

Hypothesis

Under the assumption that a generality is best that is consistent with the required terminal performance, we proposed the following hypothesis:

> Performance on test items, affect toward instruction, confidence in test item answers, and total elapsed time will be higher for students experiencing connnected generality consistency treatment than for students experiencing discrete rule consistency; and these measures for both the connected and discrete rule consistency treatments will be higher than for the content-only consistency group.

Four separate hypotheses corresponding to four dependent variables result from the above general statement. Each will be treated separately in the reporting of results.

Method

Subject Matter Content

An introductory statistics course at Brigham Young University (BYU) was selected as the experimental situation for this study. The specific matter of hypothesis testing was chosen because generally low scores from past achievement tests indicated a high level of difficulty with the topic. The subject matter met the experimental specifications of having multiple rules that could be taught as separate, discrete rules or as connected rules. The following hypothesis tests were covered in the selected unit of instruction (Christensen, 1974):

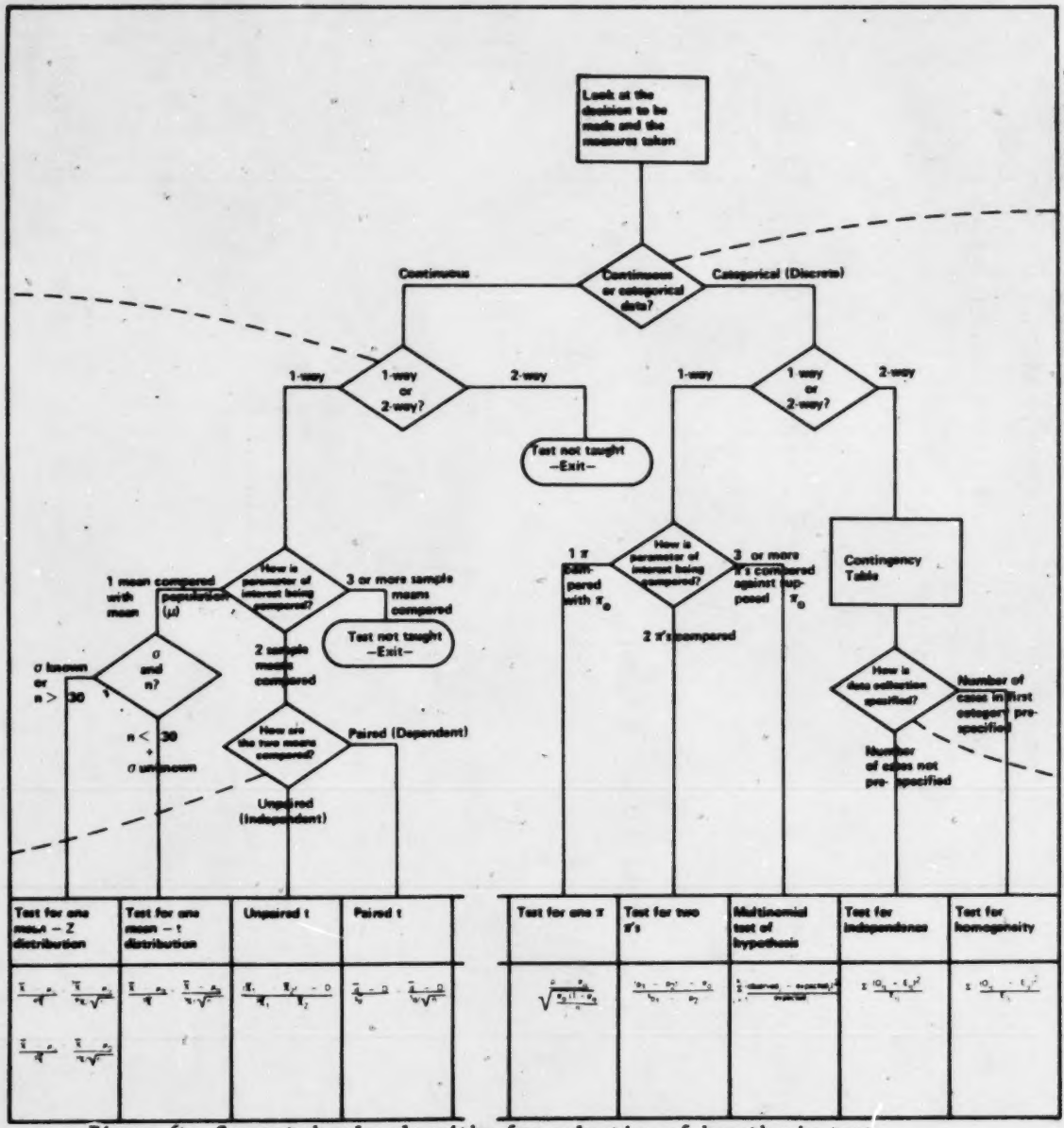
1. Test for one mean when σ is known.
2. Test for one mean when σ is not known.
3. Test for two means when the samples are independent.
4. Test for two means when observations are paired.
5. Test for one proportion.
6. Test for two proportions.
7. Chi-square test.
8. Multinomial test of hypothesis.

Subjects

The subjects were 95 regular enrollees in a college statistics undergraduate course. The course serves as one of the choices for fulfilling a general education requirement at BYU. Students received credit in the form of additional points toward the final course grade for participating in the 2-hour session and were informed that failure to participate would, in effect, penalize them, although the additional points were not dependent upon their performance.

Treatments

The study consisted of three treatments. Students in the first treatment group received a connected generality in the form of an algorithm which presented both the content operation and the task necessary to take a student from the reading of the verbally stated problem to the correct test of hypothesis and test statistics (see Figure 6). Subsequent practice provided two examples of each type of hypothesis test and test statistic. Correct answer feedback was provided on the reverse side of the practice pages. (See Figure 7 for a sample practice question.)


Look at the decision to be made and the measures taken
Continuous or categorical data?
Continuous
Categorical (Discrete)
1-way or 2-way?
1-way
2-way
Test not taught —Exit—
How is parameter of interest being compared?
1 mean compared with population mean (μ)
3 or more sample means compared
Test not taught —Exit—
σ known or n > 30
σ and n?
n < 30 + σ unknown
2 sample means compared
How are the two means compared?
Paired (Dependent)
Unpaired (Independent)
1 π compared with π0
2 π's compared
3 or more π's compared against (supposed π0
Contingency Table
How is data collection specified?
Number of cases in first category pre-specified
Number of cases not pre-specified
Test for one mean – Z distribution
Test for one mean – t distribution
Unpaired t
Paired t
Test for one π
Test for two π's
Multinomial test of hypothesis
Test for independence
Test for homogeneity

ONE-WAY CLASSIFICATION
Only one characteristic of the sample is considered:
- Example 1: Typing time.
- Example 2: Attitude toward a movie

TWO-WAY CLASSIFICATION
Two characteristics of the sample are considered at the same time:
- Example 1: Typing time and typing errors
- Example 2: Attitude toward a movie and sex of movie-goer

INDEPENDENT
The units in each sample are independent and randomly assigned.

PAIRED
1. One of a pair is assigned randomly to treatment A and one to treatment B
   —OR—
2. One unit serves as its own pair. (Such as in a pre-post situation.)

CONTINUOUS DATA
The data can be expressed in fractions. (One can weigh 165 ½ pounds.) Means and averages are usually involved. (See pp. 11-18 in text.)

CATEGORICAL (DISCRETE) DATA
Categorical data cannot be expressed in fractions. (You cannot have 1½ women.) Proportions and frequencies are usually involved. (See pp. 11-18 in text.)

INDEPENDENCE
There is no specification of the number of cases before sample is taken.

HOMOGENEITY
The number of cases in each category along one dimension is specified before sample is taken.

Figure 6. Connected-rule algorithm for selection of hypothesis test.

13. A new social studies program is supposed to produce significantly better results than a program it is to replace. Students in the course are matched on the basis of sex, IQ and G.P.A. and then the pairs are divided into "new method" and "old method" groups. Their scores on a final achievement test are taken as evidence of performance.

☐a. Select the number for the appropriate test type from list A.

☐b. Select the number for the correct test statistic from list B.

A-Type of Test	B-Test Statistic
1. test for one mean, t distribution	1. $\frac{\bar{x} - \mu_o}{s_{\bar{x}}} = \frac{\bar{x} - \mu_o}{s_x/\sqrt{n}}$
2. test for two means, dependent	2. $\frac{\bar{x} - \mu_o}{\sigma_{\bar{x}}} = \frac{\bar{x} - \nu_o}{\sigma_x/\sqrt{n}}$
3. test for one proportion	3. $\Sigma\Sigma \frac{(o_{ij} - E_{ij})^2}{E_{ij}}$
4. test for two proportions	4. $\frac{\bar{d} - D}{s_{\bar{d}}} = \frac{\bar{d} - D}{s_d/\sqrt{n}}$
5. none of the above	5. none of the above

Figure 7. A sample practice question.

Students in the second treatment condition were presented the discrete (unconnected) multiple generalities used in determining the correct test of hypothesis and test statistic. These consisted of a walk-through of separate, very simple algorithms which took the student to the appropriate test after the student's initial decision as to the specific type of data he was working with (see Figure 8). Students in this group were not given any directed strategy for connecting these behaviors or for using them as part of an overall process to help them make their initial decisions as to the nature of the statistical problem. Subsequent practice provided the student with two samples of each type of decision. Correct answer feedback was provided on the reverse side of the practice pages.

The instructional materials for the third treatment condition consisted of the regular text used in the course and directions for providing appropriate practice. The practice directions consisted of a sample item of the type used in the posttest with instructions to practice the selected items found at the end of textbook sections. No generality was provided for analyzing the type of test that a verbal practice item may pose. The text also did not help here, for each test of hypothesis was presented in a separate lesson, gave practice only in a stated kind of hypothesis, and did not request students to differentiate on the basis of kinds of test. As all students had previously been exposed to this material, this group became essentially a control group.

In the previous study (Study 1 of this report), it was hypothesized that time would decrease as a result of our treatments. In this situation, although it is desirable to reduce the amount of time students take on instruction, it is expected that time will increase, based on Bloom's (1974) observation that quality instruction initially takes longer, especially if the effect toward the instruction has on a student and confidence in mastery of the subject matter increases.[4]

Instrumentation

A brief three-question pretest was administered to get some measure of student entry behavior. The pretest was identical in form to the questions used in practice (where given) and to the posttest. Students were also asked to indicate lectures attended, materials read, and workbook practice completed in regard to the unit on hypothesis testing.

Measures for the four dependent variables of interest in the study were provided for in the treatment materials and posttest and are discussed below.

Performance. The basic performance task required the student to select (1) the appropriate hypothesis test for a problem statement, and (2) the appropriate statistical test associated with the hypothesis test (see Figure 9). The 22 multiple-choice questions provided for 44 responses. However, only the hypothesis test responses were used in the data analysis.

---

[4]Bloom does not clearly define quality instruction except for characterizing it as mastery learning.

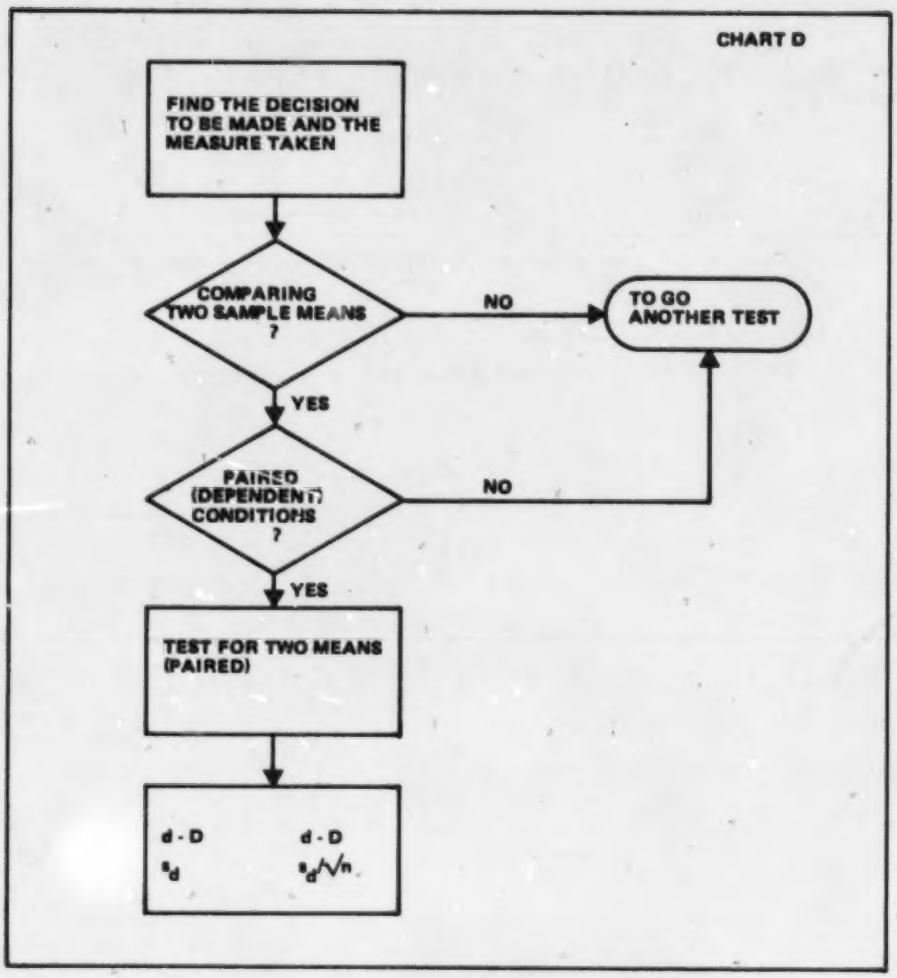


FIGURE 8. An Example of a discrete rule algorithm for selection of hypothesis test

3. Senator Incong feels that the bills introduced by members of his party in Congress will be given a positive or negative vote on the basis of party affiliation. To test this he assesses the bills over a month's time, keeping track of the party affiliation of those who took part in the voting (Democratic or Republican) and what the vote was ("For" or "Against" or "Abstain").

☐ a. Select the number for the appropriate test type from list A.

☐ b. Select the number for the correct test statistic from list B.

How confident are you in your answers to the above question?

Very Confident____ : ____ : ____ : ____ : ____ : ____ : ____ Not at all Confident

A-Type of Test

1. test for homogeneity
2. test for two means, dependent
3. test for two means, independent
4. test for two proportions
5. none of the above

B-Test Statistic

1. $\Sigma\Sigma \frac{(o_{ij} - E_{ij})^2}{E_{ij}}$
2. $\frac{(\bar{x}_1 - \bar{x}_2) - D}{s_{\bar{x}_1 - \bar{x}_2}}$
3. $\frac{\bar{d} - D}{s_{\bar{d}}} = \frac{\bar{d} - D}{s_d/\sqrt{n}}$
4. $\frac{(p_1 - p_2) - {}^{*}o}{s_{p_1 - p_2}}$
5. none of the above

Do not return to this page once you have completed your answers.

Figure 9. A sample of the basic performance task and confidence rating.

The test was designed to have more items than most students could complete in the allotted time so that differences in time and number of items completed for the separate treatment groups could be ascertained.

Affect. Questions at the end of the treatment materials and at the end of the posttest allowed students to respond to a five-category continuum of general affect in terms of how well the instruction provided preparation for performance on a test.

Confidence. A seven-point, semantic differential scale was included after each of the 22 test items in order to assess the amount of self-perceived confidence students had in their answers to the multiple-choice questions (see Figure 9).

Time. All students were to mark the time from the wall clock in a space provided at (1) the point where they finished the first 11 items on the posttest and (2) at the end of the posttest session.

Procedure

Students were randomly assigned to one of the three treatment conditions. After the pretest, students were told there would be three timed sessions, and they were requested not to begin any one of them until asked to do so. They were also informed that the materials provided would be collected before the test. Finally, they were informed that they would be provided with more items in each section than they would most likely have time to finish and that they should work steadily but carefully.

The students were given 1/2 hour for the study session. They were then requested to move on to the practice but were allowed to return to the study materials if they wished. The practice session lasted 40 minutes, after which the students recorded their sense of preparedness and attitude toward the materials used. All materials were collected. The tests were then passed out, and the students were given 30 minutes to work the 22 problems. Time was recorded on each test after the eleventh question and again at the end of the test. Students responded to a second affective measure, and the materials were collected.

Design

The three treatment groups provided three levels of the main effect, level of generality consistency. A one-way analysis of variance design provided the statistical model for both a univariate (ANOVA) and a multivariate (MANOVA) analysis of variance with two orthogonal contrasts for comparing means (control vs. the other two groups for 1 df and connected rule vs. discrete rule group for 1 df). Each of the five dependent variables was considered simultaneously in a MANOVA to correct for Type I error. ANOVA results on single dependent variables were then interpreted if the exact F-ratios from the MANOVA contrasts warranted further consideration. A generalized analysis of variance computer program which adjusted for an unbalanced design was used (Bryce & Carter, 1974).

Results

Hypothesis 1

Performance on test items for the connected rule consistency group will be higher than that for the discrete rule consistency group, and performance for both groups will be higher than the content only (low consistency) group.

Means and standard deviations on performance scores for the three treatment groups are found in Table 3. A univariate analysis of variance and an orthogonal comparison of means (Control vs. Connected and Discrete, 1 df; Connected vs. Discrete group, 1 df) indicated a significant difference between both the connected rule and discrete rule consistency groups as compared to the control group: F (2,92) = 4.21, $p < .05$. There was no significant difference between the connected rule and discrete rule consistency groups.

Table 3

Means and Standard Deviations for Dependent Variables by Treatment Group

	Treatment Group					
	Connected		Discrete			
Dependent Variable	Rule Mean	(N=33) S.D.	Rule Mean	(N=32) S.D.	Control Mean	(N=30) S.D.
Performance	10.42	.54	10.56	.54	9.10	.56
Affect Toward Instruction	3.06	.06	3.38	.07	2.92	.07
Confidence in Answers to Test Items	3.79	.23	4.19	.23	3.60	.24
Time on Posttest	24.39	.31	23.84	.32	23.9	.33

Hypothesis 2

Affect toward instruction will be higher for the connected rule consistency group than for the discrete rule consistency group, and affect for both groups will be higher than for the content only (low consistency) group.

Means and standard deviations on affect toward instruction after the treatment condition are found in Table 3. A univariate analysis of variance and an orthogonal comparison of means indicated that all three groups were significantly different from each other with the discrete rule consistency group highest, the connected rule consistency group next highest, and content only group lowest in affect: F (2,92) = 11.21, $p < .05$.

Hypothesis 3

> Confidence in answers to test items will be higher for the connected rule consistency group than for the discrete rule consistency group, and both groups will be higher in confidence than the content only (low consistency) group.

Means and standard deviations for confidence in answers to test items are found in Table 3. A univariate analysis of variance and an orthogonal comparison of means indicated no significant difference between any of the three treatment groups: F (2,92) = 1.60, $p > .05$.

Hypothesis 4

> Time to complete rule using posttest items will be longer for the connected rule consistency group than for the discrete rule consistency group, and both groups will take longer than the content only group.

Means and standard deviations for time to complete the posttest are found in Table 3. A univariate analysis of variance and an orthogonal comparison of means indicated no significant difference between any of the three treatment groups; F (2,92) = .91, $p > .05$.

Discussion

The study investigated the extent of the need for test items to be consistent with their generalities in content representation and in task behaviors on both a discrete and connected rule level. The results indicated that students who learned from materials that were consistent only on the content representation level had significantly lower scores and affect than did students whose instruction also was consistent with the test item on the task behavior level. Neither confidence nor time was significantly different across treatments.

The specific constraints of the study may have obscured greater differences, especially the hypothesized differences between the discrete rule and connected rule consistency groups. The time we could arrange demanded that we run the entire study (from introduction to instruction to practice to test) in one 2-hour sitting. Thus, the students in the connected rule consistency group, who had the most new materials to learn, had very little time to encode the rather lengthy algorithm. Given more time, we could have explicitly taught them the three or four major steps involved in the algorithm before presenting them with all the detail and, thus, allow for easier chunking (Miller, 1956) of the materials. As it was, the discrete and connected rule consistency groups may have responded more in a forward processing manner (Mayer, 1975b) in which

they did as well as they did based upon the expectations aroused, not only by the initial statement of the terminal behavior but also by the on-going practice. Familiarity also possibly impacted upon the results. The students had worked with materials similar to that provided to the discrete rule consistency group, while the algorithmic approach was not a tool familiar to the course. A study that allows adequate time for encoding of the materials, preferably run with several meetings of the groups, should be made.

Despite the constraints of the study, the presence of test-generality consistency beyond a simple content-only level clearly resulted in better test performance and affect. Teachers and developers would do well to give practice in the specific behaviors required by a terminal task.

STUDY 3: VALIDATION OF THE INSTRUCTIONAL STRATEGY DIAGNOSTIC PROFILE IN PHYSICS 100[5]

Overview and Hypotheses

The present study assumed that already designed and individualized materials on concept level tasks could be further upgraded by an example-practice-feedback sequence for each general[ity], in accordance with the hypotheses stated in Merrill and Wood (1975). The ISDP also supports the generality accepted principle that test items for classification or rule-using tasks should consist of unencountered instances. Much instruction ignores this dictum. Thus, this study analyzed the test question type for the materials used in order to compare student performance on both encountered and unencountered instance items and ascertain differences in student performance when the study materials are upgraded.

The hypotheses were based on the assumption that an increase in the degree to which tests and instruction follows the principles prescribed in the ISDP results in higher scores on tests with unencountered instance items and comparable scores on tests with previously encountered instance items.

Methods

Subject Matter Content

The subject matter consisted of six units (comprising the second quarter of the course) for an introductory physics course at Brigham Young University. There were several reasons for selecting this subject matter:

1. The course met the criteria that it be conceptually based, a quality the designers and instructors of the course desired.

2. The course, and especially the student study guide, was already carefully designed and yet showed deficiency in one or more of the areas measured by the ISDP. Too often researchers and theoreticians have been accused of shooting down straw men as they compare materials they had developed in an hypothesized better way against haphazardly presented "undesigned" lessons.

3. The materials covered a fairly broad range of topics. It is often simpler to create differences in a brief one-shot segment of material. Under such a condition, the novelty of the approach--and its brevity--go hand-in-hand to generate unusually strong attention to the task. To get at real differences, it seems necessary to have materials used over time.

4. There is a real and substantial challenge to show differences in the less neat and varied world of the on-going class rather than in the isolated laboratory setting (Glaser & Resnick, 1972). Our job is to show that the efforts which go into materials designed according to stated hypotheses result in real-life differences.

---

[5]Study conducted by M. D. Merrill, R. V. Schmidt, and R. F. Norton.

5. The course was an introductory course that serves as one of several options to fulfill a general education requirement. Thus, the students represented a broad spectrum of college undergraduates with a diversity of interests and skills.

6. Data on students enrolled in the course indicate that, during the Fall Semester of 1975, two-thirds of the students either withdrew unofficially from the course or received a grade of incomplete. Student-pacing problems obviously contributed to these results, but problems no doubt existed with the tests and instruction as well.

The material selected covered the following topics:

1. Motion and Forces.
2. Forces in Fluids at Rest.
3. Pressure in Moving Fluids.
4. Conservation of Energy.
5. Kinetic Theory of Matter.
6. Law of Increasing Entropy.

Subjects

The subjects were 43 students from two sections of a summer session of the above-mentioned course, which fulfilled part of the general education requirement in the physical sciences at the university. Other subjects, representing repeating students and students who were not present during the initial phase of the study, were too few in number within their groups to analyze meaningfully.

Treatments

The two treatments were: (1) the regularly constituted class study guide and (2) a study guide whose generalities were reinforced with example-practice-feedback segments according to ISDP principles. Moreover, eight unencountered instance item questions were added to the regular seven-item test, which consisted entirely of encountered instance or generality items. Each question had several parts to it, and there were four versions of the test.

Procedure

A preliminary study of the nature of the test question and the corresponding student performance (see the appendix) was run in order to ascertain which tests could be upgraded through eliminating previously encountered instance items and adding unencountered instance items.[6] This evaluation also allowed selection of a unit on which students showed problems in test performance. Following this, both the tests and material were upgraded according to the principles of the ISDP.

---

[6]A description of this study, which was conducted by M. D. Merrill, R. F. Norton, and R. V. Schmidt, is provided in the appendix.

Students attending class the first week were randomly assigned to the treatments. They were requested not to study with or share their materials with anyone whose materials did not match theirs. The visual difference in materials was immediately apparent. In this on-going situation, studying together had to be allowed. The randomization should have taken care of any students who might have collaborated using different materials. This was checked later through a questionnaire: two students indicated that they had looked briefly at or studied with the treatment materials they were not assigned.

Students could take the 15-item short answer essay test at a testing center at their own convenience. The experimenters picked up the test from the regular graders on a daily basis, regarded them blindly, and returned them the next day to the testing center for distribution, keeping copies of each exam for further reference.

Students not in attendance the first week (N = 6) and students retaking the course (N = 7) also took the test, but their numbers were insufficient to allow an analysis of their performances.

The amount of time required for taking the test was also recorded, and an affective questionnaire was administered after the completion of this phase of the course to see if there were any general differences between the two groups.

Design

The design was a post-test-only design with subjects nested in materials but crossed with item type. Two levels of the main effect ("regular" and "upgraded" materials) and four dependent variables (scores on encountered instance items, scores on unencountered instance items, time on test, and affect) were considered. A two-way analysis of variance design provided the statistical model for a univariable analysis of variance to test subject performance. Rummage, a generalized analysis of variance computer program to handle an unbalanced design and adjust for other effects was used in the analysis. A t-test was used to analyze time data, as not all tests carried this information.

Results

Hypothesis 1

> Tests requiring classification or rule-using behaviors for unencountered instance items result in lower scores than when the items consist of previously encountered instances.

Means and standard deviations on performance scores for the two treatment groups are found in Table 4. A univariate analysis of variance indicated a significant difference between previously encountered instance items and unencountered instance items in the hypothesized direction: $F(1,41) = 5.5$, $p < .05$.

Table 4

Mean Percentage Correct and Standard Deviations by Treatment Group

Dependent Variable	Treatment Group: Upgraded (N = 23) Mean	S.D.	Regular (N = 20) Mean	S.D.
Performance on Encountered Instance Items	76.1	02.5	74.5	02.9
Performance on Unencountered Instance Items	69.7	02.5	68.6	02.9

Hypothesis 2

Unencountered instance items requiring classification on rule-using behavior result in higher scores when the degree to which the instruction follows ISDP principles is increased over instruction which does not generally follow ISDP principles.

Means and standard deviations on performance scores for the two treatment groups are found in Table 4. A univariate analysis of variance indicated no significant difference between the two groups: $F(1,41) = .29$, $p > .05$.

Hypothesis 3

Time to complete a rule-using or classification test is greater when the degree to which instruction follows the ISDP is increased over instruction which does not generally follow ISDP principles.

As the testing place and time was out of our hands, time data was made available for only 17 of the subjects. A t-test run on the available data (Mean of 87, S.D. of 36 for the Upgraded group and Mean of 71, S.D. of 24 for the Regular group) indicated no significance: $t(17) = 1.116$, $p > .05$.

Discussion

The study attempted to ascertain if one could improve student performance on a physics test by upgrading his syllabus--mentioned earlier as the major teaching device--by adhering to ISDP principles. Because we intervened in an on-going class, this had to be attempted without controls on the teacher's lectures, videotaped helps, or the text. Though the differences noted were

in the hypothesized direction (see Table 5), they were not significant. The large standard deviations indicate that we had not captured a substantial source of variability, probably due to course materials and information beyond the syllabus. Contrary to indications on previous course participation, gleaned from several former physics students and an instructor of the course, a survey we ran indicated that all the students who answered the questionnaire (N = 28) attended virtually all the class lectures and used the text for each unit of material.

Table 5

Mean Percentage of Items Answered Correctly for Students with Upgraded and Regular Materials

	Upgraded Materials		Regular Materials	
	1st Try	2nd Try	1st Try	2nd Try
Encountered Items	76	79	76	77
Unencountered Items	69	74	66	65

Moreover, the "regular" materials were, as mentioned, already rather carefully developed. Although they did not support each stated generality directly and consistently with examples and practice, both examples and practice were available to the student who hunted for them. Thus, since we only upgraded the generalities present in the original syllabus (as a promise not to change the course for one group of students), we probably were too optimistic in the results we thought it would create. That upgrading from regular-class, "non-developed" materials does create highly significant differences has recently been demonstrated in a study comparing the results of students taught by ISDP and "regular" methods in nutrition classes (Richards, Richards, & Merrill, in press).

Our dependent measure also had constraints placed upon it which rendered it less sensitive than it could be. The instructor felt that, to keep the initial contract with his students, we had to keep the original seven questions on which the students would be graded. In order to obscure which questions these were, we had to write ours in the same essay format. Also, we were allowed only to double the length of the test, so we could not go beyond eight additional items. This was not sufficient to test all generalities at least twice, especially since we were bound to the essay form. Since greatest majority of student test items consisted of encountered instance items, and our questions consisted of unencountered instance items, this distinction was easy to break out.

Though time was not a significant effect, it would be, perhaps, given a sampling of all the students.

The sample in the affective questionnaire was too small to use for drawing any valid conclusions. There was generally mixed response from both groups, with comments, when made, indicating some dissatisfaction with the length of the upgraded materials over what they were used to but a greater security in the subsequent test answers and a desire to go to the syllabus for answers rather than moving from the syllabus to the text as some previous students indicated they did.

This research helped establish several guidelines for further intervention studies of this type. First, the experimenters should have control over all the instruction, including lecture material. They cannot assume that general nonparticipation at lectures in the past will be the case in the present. Secondly, the experimenters must have full control of the test and testing situation. This will allow a satisfactory and sensitive measure of student performance on the generalities taught as well as make possible complete data on time and affect. Following these guidelines, we can then perhaps test the power of the ISDP against materials developed at a level similar to that of the physics materials and to test ISDP developed materials over time within the framework of an on-going class. The rationale for selecting this type and amount of subject matter content, as discussed in the Methods section, is important to consider in conducting studies on the effect of instructional materials on student performance.

CONCLUSIONS

The research reviewed indicate that the propositions underlying the ISDP Profile seem to be valid. While the data reported in this document is somewhat inconclusive and not sufficient to make unqualified statements it is, nevertheless, positive. When considered with other data on the ISDP (e.g., Wood, Richards, & Merrill, 1976), it seems reasonable to assume that, when the ISDP Profile is used as a guide to analyze and modify existing instruction, the resulting performance of students is likely to be more effective. This is especially likely when the tests as well as the main line instruction can be modified. It is less likely when only the student syllabus is modified. The ISDP does seem to have considerable potential as an instructional evaluation and development tool.

RECOMMENDATIONS

1. The ISDP, as presented in the ISDP training manual, is recommended for use by Navy instructional developers and evaluators. However, it should be considered an experimental tool and should be used only by experienced instructional technologists who can appropriately adapt its use to various settings and circumstances.

2. The present effort has increased our understanding of the ISDP and has considerably increased our ability to diagnose and prescribe modifications in existing instructional materials which result in improved student performance. However, our understanding of the instructional diagnosis and prescription process has merely scratched the surface. Because of its apparent usefulness, it is recommended that ISDP validation and development efforts continue so that this instrument can become an easy to use tool for all instructional development and evaluation personnel.

REFERENCES

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Anderson, R. C., Goldburg, S. M., & Hidde, J. L. Meaningful processing of sentences. Journal of Educational Psychology, 1971, 62, 395-399.

Bloom, B. S. Time and learning. American Psychologist, 1974, 29, 682-688.

Bryce, G. R., & Carter, M. W. MAD: The analysis of variance in unbalanced designs--a software package. Presented at COMPSTAT 1974: Proceedings in Computational Statistics. Also in Bruckmann, G., Ferschl, F., & Schmetterer, L. (Eds.), Physica Verlagwien, Werzburg, Germany, 1974.

Christensen, H. B. Introductory statistics: A simplified approach. Provo, UT: Brigham Young University Press, 1974.

Ehrenpreis, W., & Scandura, J. M. The algorithmic approach to curriculum construction: A field test in mathematics. Journal of Educational Psychology, 1974, 66(4), 491-498.

Gagné, R. M. The conditions of learning. New York: Holt, Rinehart, & Winston, 1970.

Glaser, R. Components of a psychology of instruction: Toward a science of design. Review of Educational Research, 1976, 46, 1-24.

Glaser, R., & Resnick, L. B. Instructional psychology. Annual Review of Psychology, 1972, 23, 207-276.

Gropper, G. L. The design of stimulus materials in response-oriented programs. Audio Visual Communications Review, 1970, 18(2), 129-159.

Gropper, G. L. Diagnosis and revision in the development of instructional materials. Englewood Cliffs, NJ: Educational Technology Publications, 1976.

House, E. R. School evaluation: The politics and process. Journal of Educational Psychology, 1971, 62, 395-399.

Landa, L. N. Algorithmization in learning and instruction. Englewood Cliffs, NJ: Educational Technology Publications, 1974.

Markel, S. M. They teach concepts, don't they? Educational Researchers, 1975, 4, 3-9.

Mayer, R. E. Information processing variables in learning to solve problems. Review of Educational Research, 1975, 45, 525-541. (a)

Mayer, R. E. Forward transfer of different reading strategies evoked by test-like events in mathematics text. Journal of Educational Psychology, 1975, 67, No. 2, 165-169. (b)

Mechner, F. Behavioral analysis and instructional sequencing. In P. C. Lange (Ed.), Programmed Instruction. Chicago: NSSE, 1967.

Merrill, D. D., Olsen, J. B., & Coldeway, N. A. Research support for the Instructional Strategy Diagnostic Profile (Tech. Rep.). Courseware, Inc. 3 March 1976.

Merrill, M. D., & Wood, N. D. Instructional strategies: A preliminary taxonomy. Columbus, OH: Ohio State University, 1974. (ERIC Document Reproduction Service No. SE 018 771)

Merrill, M. D., & Wood, N. D. The instructional strategy diagnostic profile. Provo, UT: Courseware, Inc., 1975.

Miller, G. A. The magical number seven plus or minus two: Some limits on our capacity for processing information. Psychological Review, 1956, 63, 81-97.

Minsky, M. A framework for representing knowledge. Boston: Massachusetts Institute of Technology, Artificial Intelligence Laboratory, 1974.

Richards, S., Richards, R. E., & Merrill, M. D. Improved test performance via strategy intervention in a nutrition course. San Diego: Courseware, Inc., in press.

Scandura, J. M. Role of rules in behavior: Toward an operational definition of what (rule) is learned. Psychological Review, 1970, 77(6), 516-533.

Scandura, J. M. On higher order rules. Educational Psychologist, 1973, 10(3), 159-160.

Scandura, J. M. Role of higher order rules in problem solving. Journal of Experimental Psychology, 1974, 102(6), 984-991.

Scandura, J. M., & Durnin, J. H. Extra-scope transfer in learning mathematical strategies. Journal of Educational Psychology, 1968, 59, 350-354.

Schmidt, R. V., Wood, N. D., & Merrill, M. D. Test and generality consistency in a classification task in validation of the Instructional Strategy Diagnostic Profile (ISDP): Empirical studies (Tech. Rep.). San Diego: Courseware, Inc., in press.

Shoemaker, D. M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45, 127-147.

Stake, R. E. Measuring what learners learn. In E. R. House (Ed.). School Evaluation: The Politics and the Process. Berkely, CA: McCutchan Publishing Corporation, 1973.

Wood, N. D., Richards, R. E., & Merrill, M. D. Prediction of student performance on rule-using tasks from the diagnosis of instructional strategies. Provo, UT: Brigham Young University research paper, 1976.

Yum, K. W. An experimental test of the law of assimilation. Journal of Experimental Psychology, 1931, 14, 68-82.

APPENDIX

EVALUATION OF TEST ITEM TYPE AND STUDENT PERFORMANCE IN PHYSICS 100

Introduction

The Physics Department at Brigham Young University requested that the Division of Instructional Research, Development and Evaluation make recommendations for improving the basic physics course at the University. During the Fall Semester of 1975, two-thirds of the students either withdrew unofficially from the course or received a grade of incomplete. Student pacing problems obviously contributed to these results, but problems no doubt existed with the tests and instruction as well.

Previous evaluation of the Physics 100 course at BYU has demonstrated that the course, intended to teach conceptual matter content, contained material that was often deficient in the rule-example-practice proposition of the Instructional Strategy Diagnostic Profile.

This study examined the conceptual correspondence of the test items to the test prescriptions of the ISDP. Student performance was compared on the various types of items that were included on the tests.

Method

The Physics Department provided the pool of test items from which all of the tests administered to the students were constructed. Each of the test items was classified into one of five categories according to the type of content they measured:

1. Unencountered inquisitory instances (Ieg)--for questions in which the student was asked to apply a rule (given or not given) to a particular instance not previously encountered.

2. Partially encountered inquisitory instance--for the same type of questions as on number one above, but where the particular instance had been only partially encountered before.

3. Encountered inquisitory instance--for the same type of questions as in numbers one and two above, but where the particular instance had been previously encountered in the instructional materials.

4. Inquisitory generality (IG)--for questions in which the student was asked to remember or recognize a rule statement or concept definition.

5. Miscellaneous category (M)--for questions where the student was asked: (a) to cite evidence (data or logic) for a given proposition, (b) to give or recognize superordinate, coordinate, or subordinate relationships among or between propositions or concepts, or (c) to remember a given constant or some specific piece of data, a fact, etc., which is an identity.

Most of the test items contained more than one category of question within the item. If any Ieg questions occurred within an item, the whole item was classified Ieg. If an IG question was combined with a M question, the whole item was classified M. Interrater reliability was strengthened by having both raters rate the same items separately and then compare the results. The few disagreements were discussed until consensus was reached on all items.

An item was classified as encountered if the answers to two-thirds or more of the questions constituting the item were found anywhere in the text, the syllabus, or the television lectures. An item was classified as unencountered if one-third or fewer of the questions constituting the item were encountered in the above mentioned sources. Items falling in between these two cutoff points or items where a similar but not identical instance was encountered in the lesson materials were classified as partially encountered.

Independent Variables

An analysis of variance was run, using three independent variables: (1) the three examinations over three different areas of subject matter, (2) the seven test items used on each test, and (3) the five categories indicating the content type of each item.

The first examination covered the first six chapters of the text and aimed at a conceptual understanding of Newton's first two laws of motion. The second examination covered chapters seven through ten of the text and aimed at a conceptual understanding of the laws of force and motion, conservation of energy, the kinetic theory of matter, and the law of entropy. The third examination covered chapters 11 through 14 and aimed at a conceptual understanding of the properties of waves, electricity, and magnetism.

The test item number was included as an independent variable because it served as an index of the difficulty level of the various items. The sixth and seventh items on each test (A level items) were designed by the developers of the test to be more difficult than the fourth and fifth items (B level items), and these in turn were designed to be more difficult than items one, two, and three (C level items). An inclusion of this variable in the analysis of variance enabled an empirical evaluation of the preassessed difficulty levels of the items.

The five content type categories were included to assess which types of questions were being answered most effectively by the students.

Dependent Variables

The Physics Department had already gathered data on the number of students that had missed each item in the pool of test items. The number of times each item was used on a test was calculable from knowing the total number of tests given and the procedure used to generate the various tests that were used. From the above information, the percentage of students answering each test item correctly could be determined. This percentage was used as the dependent variable in the analysis of variance reported in the results section of this paper.

The final data analysis design was a 3 x 7 x 5 matrix that can most clearly be understood by looking at the design diagram in Figure A-1.

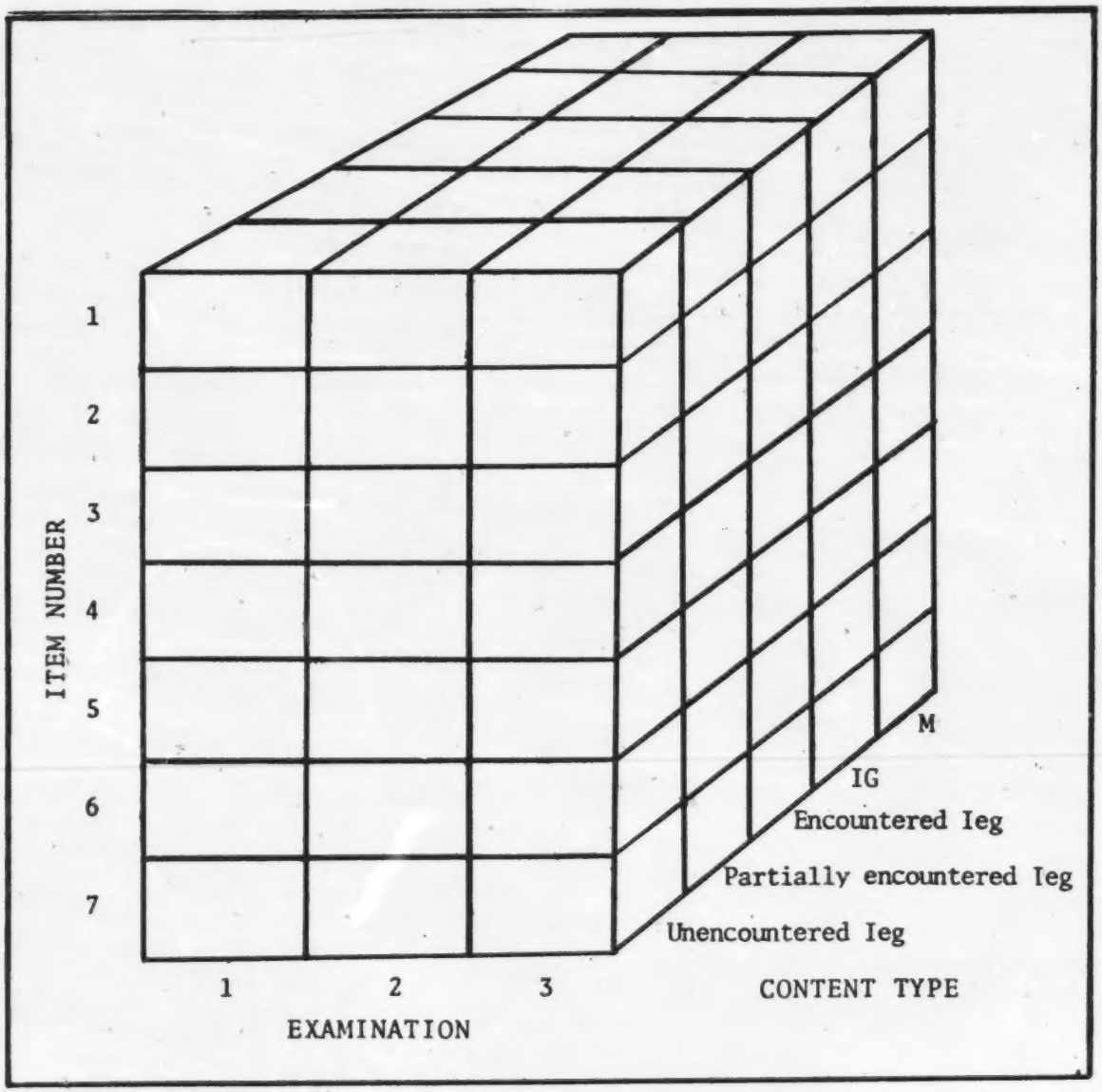


Figure A-1. Design for evaluation of test item type and student performance.

Results

Table A-1 shows the mean percentages of students answering items correctly on each of the three examinations, the mean percentages of students answering questions correctly under each test item number, and the means formed by the examination and item number interactions. An analysis of variance showed that the differences among examinations means were significant ($F = 16.366$, $p < .01$) as were differences among item number means ($F = 23.747$, $p < .01$).

Table A-1

Examination vs. Item Number Matrix Mean Percentage of Students Answering Items Correctly

Item Number	Examination 1	2	3	Means
1	87.945	83.141	82.998	84.816
2	86.447	80.196	75.807	81.250
3	72.444	82.376	72.177	75.418
4	79.284	71.663	61.232	71.590
5	73.763	63.785	67.017	68.745
6	72.300	60.378	46.421	58.372
7	58.679	60.142	56.509	58.182
Means	76.021	73.291	65.811	

Table A-2 shows the mean percentages of students answering each type of test question correctly, the mean percentages of students answering questions correctly under each test item number, and the means formed by the content type and item number interactions.

Table A-3 shows the mean percentages of students answering items correctly on each of the three examinations, the mean percentages of students answering each type of test question correctly, and the means formed by examination and content type interactions. An analysis of variance showed that the differences among the content-type means (66.077, 72.505, 73.388, 73.024, and 74.159) were also significant ($F = 2.649$, $p < .05$).

The nature of these differences was analyzed using prediction coefficients and is reported in the discussion section which follows. Table A-4 gives the percentages of items used from each content type on each examination.

Table A-2

Content Type vs. Item Number Matrix Mean Percentage of Students Answering Items Correctly

Item Number	Question Type					Means
	Unencountered Ieg	Partially Encountered Ieg	Encountered Ieg	IG	M	
1	82.045	86.137	87.556	80.146	87.273	84.816
2	77.138	83.243	78.217	87.611	79.333	81.250
3	59.209	79.713	74.180	78.796	81.220	75.418
4	61.622	70.549	77.684	72.476	72.666	71.590
5	64.186	92.000	70.486	63.813	0	68.745
6	58.785	45.253	66.425	41.333	76.250	58.372
7	64.148	53.574	59.166	65.864	48.610	58.182
Means	66.077	72.505	73.388	73.024	74.159	

Table A-3

Examination vs. Content Type Mean Percentage of Students Answering Correctly Items of Each Question Type

Question Type	Examination			Means
	1	2	3	
Unencountered Ieg	71.876	64.144	61.568	66.077
Partially Encountered Ieg	78.058	73.504	59.735	72.505
Encountered Ieg	78.084	75.238	66.842	73.388
IG	78.200	73.470	68.204	73.024
M	73.544	78.495	70.743	74.159
Means	76.021	73.291	65.811	

Table A-4

The Percentages of Items Used from Each Question Type on Each Examination

Question Type	Examination 1	2	3
Unencountered Ieg	31%	20%	11%
Partially Encountered Ieg	19%	7%	5%
Encountered Ieg	25%	35%	46%
IG	10%	34%	29%
M	15%	4%	9%
	100%	100%	100%

Discussion

The difference among examinations indicates that significantly fewer students responded correctly to the items on the third exam (see Table A-1). It is possible that the items were more difficult or that, because of the end of the semester, fewer students retook the third exam than retook the first and second exams. Each time a student retook an examination, even though the items were different than on the previous exam, he was likely to do better than he did the time before because of more study in the area where he was deficient. This would mean that the average percentage of students answering items correctly was artificially elevated for both the first and second exams--more so for the first than for the second. Regardless of the question type, items on the third exam were missed more often than the corresponding type of items on the other two exams (see Table A-3).

The sixth and seventh test questions on each exam were consistently more difficult than all of the other questions (see Table A-1). However, question six on exam one was not significantly more difficult than questions three and five. The overall means for the seven question numbers indicate that questions four and five fell in the middle range of difficulty as intended, but this was not consistent when the three exams were considered separately.

Although questions six and seven were more difficult than the others, they were not measuring a higher level of conceptual understanding, as might be hoped, but, rather more obscure details encountered in the test, syllabus, or videotapes. It might be more meaningful to use previously unencountered questions as A and B level items. This would tend to award Bs and As on the basis of a better conceptual understanding of the material rather than on the basis of ability to remember more obscure detail.

Regardless of the type of question involved, items six and seven were consistently missed more frequently than the other item (see Table A-2). This is likely a reflection of the tendency for A-level items to deal with obscure details. It is also interesting that on the unencountered inquisitory instance questions (unencountered Ieg) for the B and C levels, each of the mean percentages falls below the grand mean for its respective question number. This is as we would expect for more difficult questions. Yet for the A level questions, the unencountered Ieg questions have mean percentages equal to or higher than the grand means for their respective question numbers. This may mean that the students have acquired a set response to unencountered unobscure items versus unencountered obscure items. For example, they may be skipping the unencountered unobscure items without spending much time on them because they realize that they have never seen them before. At the same time, because the A level items involve more obscure material, they are spending more time to think and are coming up with more right answers on their own.

The unencountered instance questions are significantly more difficult ($p < .05$) than all other test question types as we might expect if they were measuring understanding at conceptual level rather than at just a memory level (see Table A-3). The partially encountered instance questions, the encountered instance questions, the inquisitory generality questions, and the miscellaneous questions all seem to be at the same level of difficulty for the students. However, the partially encountered instance items on test three are slightly (though not significantly) more difficult than the unencountered instance items. The partially encountered items in tests one and two, but more similar to the unencountered items in test three.

The percentages of unencountered instance questions on the various exams are also very interesting. Because they are the most difficult items, one might have expected a positive correlation between the percentages of such items and the performance by the students on the tests. However, there was a negative correlation (see Table A-4). Although test three was the most difficult for the students, it had only 11 percent of the most difficult question type. This might mean that the subject matter tested in test three was inherently more difficult or that the instruction in this area was weaker.

The present study will be expanded to see if student performance on unencountered instance questions could be improved by following the princples of effective instruction recommended in the Instructional Strategy Diagnostic Profile.

DISTRIBUTION LIST

Chief of Naval Operations (OP-987P10), (OP-991B)
Chief of Naval Education and Training (00A)
Chief of Naval Education and Training Support
Chief of Naval Education and Training Support (01A), (N-5)
Chief of Naval Technical Training (Code 016)
Chief of Naval Material (NMAT 035)
Chief of Naval Research (Code 450) (4)
Chief of Naval Personnel (pers-10c)
Chief of Information (OI-2252)
Commanding Officer, Naval Aerospace Medical Institute (Library Code 12) (2)
Commanding Officer, Naval Education and Training Program Development Center
Commanding Officer, Naval Development and Training Center (Code 0120)
Officer in Charge, Naval Education and Training Information Systems Activity
Director, Training Analysis and Evaluation Group (TAEG)
Director, Defense Activity for Non-Traditional Education Support
Personnel Research Division, Air Force Human Resources Laboratory (AFSC)
Lackland Air Force Base
Occupational and Manpower Research Division, Air Force Human Resources
Laboratory (AFSC), Lackland Air Force Base
Technical Library, Air Force Human Resources Laboratory, Lackland Air Force Base
Technical Training Division, Air Force Human Resources Laboratory,
Lowry Air Force Base
Program Manager, Life Science Directorate, Air Force Office of Scientific
Research (AFSC)
Army Research Institute for the Behavioral and Social Sciences
Coast Guard Headquarters (G-P-1/62)
Military Assistant for Training and Personnel Technology, ADDR&E, OAD(E&LS)
Director for Acquisition Planning OASD(I&L)
Defense Documentation Center (12)